Matching in the Dark: A Dataset for Matching Image Pairs of Low-light Scenes

Wenzheng Song[1]   Masanori Suganuma[1,2]   Xing Liu[1]
Noriyuki Shimobayashi[1]   Daisuke Maruta[3]   Takayuki Okatani[1,2]

[1]GSIS, Tohoku University   [2]RIKEN Center for AIP   [3]Socionext Inc.

{song,suganuma,ryu,nshimobayashi,okatani}@vision.is.tohoku.ac.jp

maruta.daisuke@socionext.com

Abstract

This paper considers matching images of low-light scenes, aiming to widen the frontier of SfM and visual SLAM applications. Recent image sensors can record the brightness of scenes with more than eight-bit precision, available in their RAW-format image. We are interested in making full use of such high-precision information to match extremely low-light scene images that conventional methods cannot handle. For extreme low-light scenes, even if some of their brightness information exists in the RAW format images' low bits, the standard raw image processing on cameras fails to utilize them properly. As was recently shown by Chen et al.[14], CNNs can learn to produce images with a natural appearance from such RAW-format images. To consider if and how well we can utilize such information stored in RAW-format images for image matching, we have created a new dataset named MID (matching in the dark). Using it, we experimentally evaluated combinations of eight image-enhancing methods and eleven image matching methods consisting of classical/neural local descriptors and classical/neural initial point-matching methods. The results show the advantage of using the RAW-format images and the strengths and weaknesses of the above component methods. They also imply there is room for further research.

1. Introduction

Structure-from-motion (SfM) [24, 58] and visual SLAM (simultaneous localization and mapping) [38, 53] have been used for real-world applications for a while. The mainstream methods use point correspondences between multiple views of a scene. They first detect keypoints and extract the descriptor of the local feature at each keypoint [31, 35, 2, 44]. They then find initial point correspondences between images and eliminate outliers from them, finally estimating the geometric parameters such as camera poses, etc.

SfM and visual SLAM have the potential to widen their application fields. One important target is the application to extremely low-light environments, such as outdoor scenes at night under moonlight or indoor scenes with insufficient illumination. Making it possible to use SfM and visual SLAM in these environments is essential for real-world applications, such as autonomous vehicles that can operate at night.

Owing to the advancement of image sensors, they can record incoming light with more than eight bits (e.g., 14 bits). However, standard raw image processing employed on many cameras cannot make full use of the information existing in the lower bits of the sensor signal; it reduces mosaic artifacts on the sensor signal, adjusts the white balance and contrast, and then converts the processed signal into the standard format of eight-bit RGB images (we will refer to this raw image processing as RIP in this paper). This limitation arguably comes from the requirement for versatility against all sorts of scenes with various lighting conditions in addition to reducing the number of bits. In extreme low-light scenes, even when some details of the scenes' brightness are stored in the low bits of their RAW signals, the standard RIP often yields mostly black images. The study of SID (see-in-the-dark) [14] well proves such limitation of the image pipeline, in which the authors show that a CNN can learn to convert such RAW-format images of dark scenes into brightened images with a natural appearance.

It is very likely that we can do the same with SfM and visual SLAM applied to low-light scenes, i.e., extracting the information present in the lower bits of the RAW signals to make SfM/visual SLAM work. The question is how to do this. It is noteworthy that the goal is not to generate natural looking bright images as SID does but to achieve the optimal performance for SfM and visual SLAM.

There are potentially several directions to achieve the goal. One is to develop a keypoint detector and a feature descriptor that work directly on the RAW-format images. Even if keypoint detectors and descriptors are not good

enough, it could be possible to attain the necessary level of matching performance by strengthening the subsequent steps in the pipeline. Recently, CNNs have been applied to these steps, leading to promising results, such as outlier removal in the initial correspondences [37, 12] and establishing initial matching [45]. In parallel to these, the application of image enhancement methods for RAW-format low-light images in a pre-processing stage of image matching could be useful, e.g., SID [14] and others [13, 57]. Methods for more general image restoration would be applied to the RAW-format images [63, 29].

As above, we can think of multiple different approaches to making SfM and visual SLAM methods applicable to low light environments. To promote further studies, we need a dataset to evaluate the above approaches in a multi-faceted fashion. Aiming at widening their application field toward lower-light scenes, it is necessary to examine how underexposed the image will be that each approach can deal with. There is currently no dataset that can be used for this purpose. Considering these, we create a dataset having the following features:

- To examine each method's limit with underexposed images, we acquire multiple RAW-format images at each scene's position with $48 = (6$ shutter speeds $\times$ 8 ISO settings) exposure settings ranging from extreme to mildly underexposure settings. The camera is mounted on a tripod while capturing all the images.

- We additionally provide long-exposure images, using which as the ground truth, one can evaluate image restoration methods on the task of estimating it from one of the underexposed images.

- The current standard for the evaluation of image matching methods is to measure the accuracy of the downstream task, i.e., the estimation of geometric parameters, as was pointed out in recent studies [27]. Therefore, we acquire images from two positions to form stereo pairs for each scene along with their ground truth relative pose. To obtain the ground truth pose, we capture a good quality image with a long-exposure setting for each scene position.

- The dataset contains diverse scenes consisting of 54 outdoor and 54 indoor scenes.

Using this dataset, we experimentally evaluate several existing component methods for the SfM pipeline, i.e., detecting keypoints and extracting descriptors [31], finding initial point correspondences, and removing outliers from them [20, 12, 45]. We choose classical methods and learning-based methods for each. We also evaluate the effectiveness of image enhancement, including classical image-enhancing methods with/without denoising [16],

and a CNN-based method [14, 62]. The results show the importance of using the RAW-format images instead of using the processed images by the standard RIP. They further provide the strengths and weaknesses of the above component methods, also showing that there is room for further improvement.

2. Related Work

2.1. Matching Multi-view Images

Matching multi-view images of a scene is a fundamental task of computer vision, and its research has a long history. It generally performs the following steps: detecting keypoints/computing local descriptors, establishing initial point correspondences, and removing outliers to finding correct correspondences. A baseline of this pipeline, built upon traditional methods, consists of SIFT [31], SURF [8], etc. for detecting interest points and extracting their local descriptor, nearest neighbor search in the descriptor's space for obtaining initial correspondences across images, with an optional 'ratio test' step for filtering out unreliable matches [31], and RANSAC for outlier removal [20, 43].

A recent trend is to use CNNs to detect keypoints and/or extract local descriptors. Early studies attempted to learn either keypoint detectors [56, 48, 7], or descriptors [51, 61, 23, 5, 60]. In contrast, in recent studies, researchers have proposed end-to-end pipelines [59, 17, 19, 40, 26, 39] that can perform the two at once. Despite the success of CNNs in many computer vision tasks, it remains unclear that these learning-based methods have surpassed the classical hand-crafted methods. In parallel to the developments of methods for keypoint detectors and descriptors, several recent studies have developed learning-based methods for initial point matching and outlier removal [12, 45].

2.2. Datasets for Image Matching

There are many datasets created for the research of image matching [36, 1, 64, 42, 52, 28, 56]. Many recent studies of image matching employ HPatches [6]. There are also a number of datasets for visual SLAM and localization/navigation [47, 21, 33, 46, 4].

Some of these datasets provide challenging cases, including illumination changes, matching daylight and nighttime images, motion blur in low-light conditions, etc. However, all these datasets provide only images in the regime where the standard RIP can successfully yield RGB images with a well-balanced brightness histogram. This is also the case with a recent study [25] that analyzes image retrieval under varying illumination conditions. Our dataset contains the images of very dark scenes all in a RAW-format with 14-bit depth. In fact, while we have verified the authors' findings in [25] with 8-bit images converted from our RAW-format images using the standard RIP, they do not hold in

the case of directly using the RAW-format images, as we will show later.

There are also many evaluation methods for image matching, which are developed aiming at a more precise evaluation [49, 36, 1, 15, 10]. A recent study has introduced a comprehensive benchmark for image matching [27]. As in this study, the current trend is to focus on the downstream task; the accuracy of the reconstructed camera pose is chosen as a primary metric for evaluation. Following this trend, our dataset provides the ground truth for the relative camera pose between every stereo image pair.

2.3. Image Enhancement

There are many image-enhancing methods that improve the quality of underexposed images. Besides basic image processing such as histogram equalization, there are many methods based on different assumptions and physics-based models, etc., such as global analysis and processing based on the inverse dark channel prior [34, 18], the wavelet transform [32], the Retinex model [41], and illumination map estimation [22]. These methods are proven to be effective for images that are mildly underexposed.

To deal with more severely underexposed images, Chen *et al.* proposed a learning-based method that uses a CNN to directly convert a low-light RAW image to a good quality RGB image [14]. Creating a dataset containing pairs of underexposed and well-exposed RAW images (i.e., the SID dataset), they train the CNN in a supervised fashion. Their method can handle more severe image noise and color distortion emerging in underexposed images than the previous methods. For the problem of enhancing extreme low-light videos, Chen *et al.* extended this method while creating a dataset for training [13]. In parallel to these studies, Wei *et al.* have developed a model of image noises, making it possible to synthesize realistic underexposed images [57]. They demonstrated that a CNN trained on the synthetic dataset generated by their model performs denoising equally well or even better than a CNN trained on pairs of real under/well-exposed images.

While these studies aim sorely at image enhancement, our study considers the problem of matching images of extremely low-light scenes. Our dataset contains stereo image pairs of multiple scenes; there are 48 low-light RAW images with different exposure settings and one long-exposure reference image for each camera position of each scene. It is noteworthy that they include much more underexposed images than the datasets of [14, 13].

3. Dataset for Low-light Image Matching

3.1. Design of the Dataset

We built a dataset of stereo images of low-light scenes and named it the MID (Matching In the Dark) dataset. It
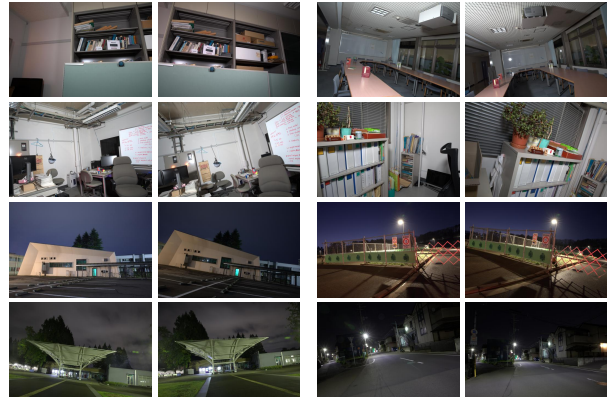

Figure 1. Example stereo image pairs (long exposure versions) of four indoor scenes (upper two rows) and four outdoor scenes (lower two rows).

contains stereo image pairs of 54 indoor and 54 outdoor scenes (108 in total). We used a high-end digital camera to capture all the images; they are recorded in a RAW format with 14-bit depths. Figure 1 shows example scene images. For each of the 108 scenes, we captured images from two viewpoints with 49 different exposure settings, i.e., 48 exposure settings in a fixed range plus one long exposure setting to acquire a reference image. Note that most of the images are so underexposed that the standard RIP cannot yield reasonable RGB images from them.

Some of the 48 images of each scene captured with the most underexposed settings are so underexposed that they appear to store only noises; it will be impossible to perform image matching using them, even if we try every one of the currently available methods. Nevertheless, we keep these images in the dataset to assess the lower limit of exposure to which image matching and restoration methods work, not only existing methods but those to be developed in the future. We designed the dataset primarily to evaluate image matching methods in low-light conditions, but the users can also evaluate image-enhancing methods. Our 48 images of each scene contain more severely underexposed ones than any existing datasets for low-light image enhancement (e.g., [14]).

3.2. Detailed Specifications

The dataset contains $10,584 (= 108(\text{scenes}) \times 2(\text{stereo}) \times (48 + 1)(\text{exposure settings}))$ images in total. They are of $6,720 \times 4,480$ pixels and in a RAW format of 14 bits per pixel; its Bayer pattern is RGGB. We used Canon EOS 5D Mark IV with a full-frame CMOS sensor and EF24-70mm f/2.8L II USM to capture these images.

For each scene, we set up the camera in two positions to capture stereo images. For each position, we mounted the camera on a sturdy tripod while capturing 49 images. We first captured a long-exposure image, which serves as a ref-

erence image; we use it to compute the ground-truth camera poses of the stereo pair, as will be explained in Sec. 3.3. To capture the reference image, we choose exposure time from the range of 10 to 30 seconds, while fixing ISO to 400.

We then captured the low-light images in 48 different exposure settings that are combinations between six exposure times and eight ISO values. The exposure time is chosen from the range of $[1/200, 1]$ seconds for the indoor scenes and $[1/200, 0.5]$ seconds for the outdoor scenes. The ISO value is chosen from $\{100, 200, 400, 800, 1600, 3200, 6400, 12800\}$.

The indoor scene images were captured in closed rooms with regular lights turned off; the illuminance at the camera is in the range of 0.02 to 0.3 lux. The outdoor scene images were captured at night under moonlight or street lighting. The illuminance at the camera is in the range of 0.01 to 3 lux.

3.3. Obtaining Ground Truth Camera Pose

To compare various image matching methods with different local descriptors and keypoint detectors, we need to evaluate the accuracy of the camera poses estimated from their matching results. We consider stereo matching in our dataset, and an image matching method yields the estimate of the relative camera pose between the stereo images. To obtain its ground truth, we use the pairs of the reference images to perform image matching, from which we estimate the relative camera pose for each scene. Following [11], we use it as the ground truth after manual inspection along with correction, if necessary, which we will explain later.

The detailed procedure for obtaining the ground truth camera pose for each scene is as follows.

We first convert the two reference images in the RAW format into RGB space[1]. We then covert each RGB image into grayscale and compute keypoints and their descriptors using the difference of Gaussian (DoG) operator and the RootSIFT descriptor [3]. We next establish their initial point matches using nearest neighbor search with Lowe's ratio test [31] with a threshold of 0.8.

We then estimate the essential matrix by using the 5-point algorithm with the pretrained neural-guided RANSAC (NG-RANSAC) [12]. We employed the authors' implementation for it. We employ NG-RANSAC over conventional RANSAC, since we found in our experiments that it consistently yields more accurate results. Calibrating the camera with the standard method using a planar calibration chart, we decompose the estimated essential matrix and obtain the relative camera pose (i.e., translation and rotation) between the stereo pair.

As mentioned above, we performed a manual inspection

---

[1]Following [14], we used rawpy (https://pypi.org/project/rawpy/), a python wrapper for *libraw* that is a raw image processing library (https://www.libraw.org/)

of the estimated essential matrix, ensuring they are reliable enough to be used as the ground truths. We did this by checking if any image point on the paired images satisfies the epipolar constraint given by the estimated essential matrix. To be specific, we manually select a point on either left or right image and draw its epipolar line on the other image.

We then visually check if the corresponding point lies on the epipolar line with a deviation less than one pixel. We chose a variety of points having different depths for this check. If an estimated essential matrix fails to pass this test, we either remove the scene entirely or manually add several point matches to get a more accurate estimate of the essential matrix and perform the above test again. All the scenes in our dataset have passed this test.

4. Matching Images in Low-light Scenes

This section discusses what methods are applicable to matching low-light images in our dataset. We evaluate those in our experiments.

4.1. Conversion of RAW Images to RGB

As there is currently no image matching method directly applicable to RAW-format images, we consider existing keypoint detectors and local descriptors that receive grayscale images. To cope with the low-light condition, we plug image enhancement methods before keypoint detectors and local descriptors, which we will describe later.

It is first necessary to convert RAW-format images into RGB/grayscale images. We have two choices here. One is to use the standard RIP that converts RAW to RGB. As mentioned in Sec. 1, the standard RIP often fails to make use of brightness information stored in the lower bits of RAW signals of dark scenes, due to the requirement for versatility against a variety of scenes with different illumination conditions and also the limit of computational resources available to on-board RIP. To confirm its limitation, we evaluate this standard-camera-pipeline-based conversion in our experiment; we use the LibRaw library using rawpy, a Python image processing module.

The other choice is to do the conversion without using the standard RIP. We will explain this below, because it is coupled with the image enhancement step.

4.2. Image Enhancement

Thus, we consider two methods, i.e., using the standard RIP for the RAW-to-RGB conversion and directly using RAW-format images. For each, we consider three different image enhancing methods.

4.2.1 Conversion by standard camera pipeline

When using the standard RIP to convert RAW images, we consider applying the following four methods to its out-
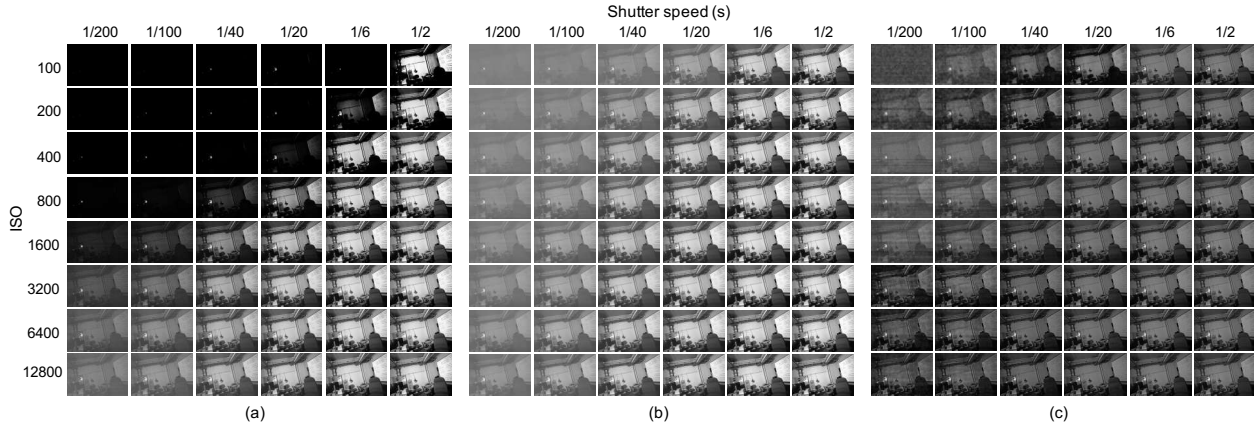
Figure 2. Images of a scene captured from the same camera pose that are converted from their RAW-format originals by three conversion methods. (a) **RIP-HistEq**. (b) **Direct-BM3D**. (c) **SID**. See text for these methods.
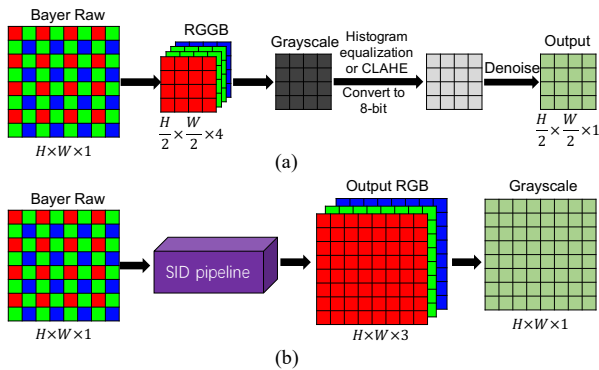

Figure 3. The pipelines of two image-enhancing methods. (a) **Direct-HistEq** or **Direct-CLAHE**. (b) **SID**.

puts: none, a classical histogram equalization, a contrast limited adaptive histogram equalization (CLAHE), and a CNN-based image enhancement, MIRNet [62]. We choose MIRNet because it is currently the best image-enhancing method applicable to RGB/grayscale images. Figure 2(a) shows examples of the standard RIP with histogram equalization. We will refer to the four methods as **standard RIP**, **RIP-HistEq**, **RIP-CLAHE**, and **RIP-MIRNet** in Sec. 5.

4.2.2 Direct Use of RAW-format Images

We consider two approaches. One is to use standard image processing methods to convert RAW to RGB/grayscale images; see Fig. 3(a). For this, we employ the following simple approach. Given a Bayer array containing the input RAW data, we first apply black level subtraction to it and then split the result into four channels; the pixel values are now represented as floating point numbers. We then take the average of the two green channels to obtain an RGB image and convert it to grayscale using the OpenCV function

`cvtColor`. Next, we perform histogram equalization or CLAHE to improve the brightness of the image. We map the brightness in the range $[m - 2d, m + 2d]$, where m is the average brightness and d is the mean absolute difference from m to each pixel value, to the range $[0, 255]$. Finally, we quantize the pixel depth to 8 bits. We will call this method **Direct-HistEq** or **Direct-CLAHE**.

We optionally apply denoising to the converted image at the final step. We employ BM3D [16] with a noise PSD ratio of 0.08 in our experiments. The resulting image will be transferred to the second step of image matching. Figure 2(b) shows examples of the converted images by the method. We will call this method **Direct-BM3D**.

In parallel to the above, we consider a CNN-based image enhancing method that directly works on RAW-format images; see Fig. 3(b). We employ SID [14], a CNN trained on the task of converting an underexposed RAW image of a low-light scene to a good quality image. It is designed to receive the RAW data of an image and output an RGB image. We calculate the amplification ratio of SID using shutter speed and ISO values between the underexposed and the reference images. As the output of SID is twice as large as others, we downscale the image size by $2 : 1$ and then convert it into grayscale for image matching; see Fig. 2(c). We used the pretrained model provided by the authors , which is trained on the SID dataset. We call this method **SID** in what follows.

4.3. Image Matching

We consider matching a pair of images of a scene here. It is to establish point correspondences between images while imposing the epipolar constraint on them and estimate the camera pose (*i.e.*, a essential or fundamental matrix) encoded in the constraint. The standard approach to the problem is to first extract keypoints and their local descriptors from each input image, establish initial matching of the key-

points between the images, and finally estimate the camera pose from them.

There are at least several methods for each of the three steps. There are many classical methods that do not rely on learning data. As with other computer vision problems, neural networks have been applied to each step. They were first applied to the first step, *i.e.*, keypoint detectors [56] and descriptors [23, 51, 5, 60], to name a few. The next was the third step of robust estimation [37, 12]. Recently, SuperGlue [45] was proposed, which deals with the step of establishing initial point correspondences.

5. Experiments

We experimentally evaluate the combinations of several methods discussed in Sec. 4 using our dataset.

5.1. Experimental Configuration

5.1.1 Compared Methods

We choose both classical methods and neural network-based methods for each step of image matching. As for keypoint detection and local descriptors, we choose Root-SIFT [3] and ORB [44] as representative classical methods; we consider ORB because it has been widely used for visual SLAM. We use their implementation of OpenCV-3.4.2. We use SuperPoint [17], Reinforced SuperPoint [9], GIFT [30], R2D2 [26], and RF-Net [50] as representative neural network-based methods. Furthermore, we employ L2-Net [54] and SOSNet [55] as hybrid methods of classical and neural-based methods; they compute local descriptors based on the SIFT keypoints and neural networks. For them, we use the authors' implementation and follow the settings recommended in their papers.

As for outlier removal of point correspondences, we choose RANSAC and NG-RANSAC [12]. We use the OpenCV-3.4.2 implementation of RANSAC with `threshold = 0.001`, `probability = 0.999`, and `maxIters = 10,000` with the five point algorithm and use the authors' code for the latter. For obtaining initial point correspondences, we use the nearest neighbor search and also SuperGlue [45]. We apply Lowe's ratio test [31] with a threshold of 0.8 to RootSIFT, L2-Net, SOSNet, and RF-Net.

To summarize, we compare the following eleven methods: **SP:** Superpoint + NN + RANSAC, **RSP:** Reinforced SuperPoint + NN + RANSAC, **GIFT:** GIFT + NN + RANSAC, **SP + SG:** SuperPoint + SuperGlue + RANSAC, **R2D2:** R2D2 + NN + RANSAC, **RF:** RF-Net + NN + RANSAC, **L2:** L2-Net + NN + RANSAC, **SOS:** SOSNet + NN + RANSAC, **RS:** RootSIFT + NN + RANSAC, **RS + NG:** RootSIFT + NN + NG-RANSAC, **ORB:** ORB + NN + RANSAC. As for image enhancers, we use the eight methods explained in Sec. 4.2., i.e.,
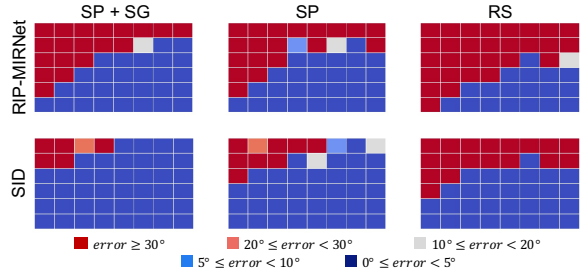

Figure 4. Angular errors of the camera pose estimated by several methods for a scene from images with 6×8 different exposure settings. The number of cells with an error lower than a specified threshold quantifies the robustness of the method.

standard **RIP**, **RIP-HistEq**, **RIP-CLAHE**, **RIP-MIRNet**, **Direct-HistEq**, **Direct-CLAHE**, **Direct-BM3D**, and **SID**. We combine these eight image enhancers with the above eleven image matching methods and evaluate each of the 88 pairs. We resize the output images from each image enhancer to 960×640 pixels and feed it to the image matching step.

5.1.2 Evaluation

We compare these methods by evaluating the accuracy of their estimated relative camera pose. We apply each pair of an image enhancer and an image matching method to the stereo images of each scene. We consider only pairs of stereo images with the same exposure setting; there are 48 pairs per each scene. Thus, we have 48 estimates of relative camera pose for each scene.

To evaluate the accuracy of these estimates, we follow the previous work [37, 12, 45]. Specifically, we measure the difference between the rotational component of the ground truth camera pose and its estimate, as well as the angular difference between their translational components. We use the maximum of the two values as the final angular error. Figure 4 shows examples of the results. Each of the colored 6×8 matrices indicate the above angular errors of one of the compared methods for a scene and the 48 exposure settings.

We are interested in how robust each method will be for underexposed images. To measure this, we count the exposure settings (out of 48) for which each method performs well. To be specific, denoting the above angular error of i-th exposure setting by e_i ($i = 1, \ldots, 48$), we set a threshold τ and count the exposure settings with an error lower than τ as $N_\tau = \sum_{i=1}^{48} 1(e_i < \tau)$, where $1(\texttt{True}) = 1$ and $1(\texttt{False}) = 0$. We normalize N_τ dividing by the total number of exposure settings. As shown in Fig. 4, the angular error decreases roughly in a monotonic manner from well-exposed toward underexposed settings. Thus, a larger N_τ means that the method is more robust to underexposure.
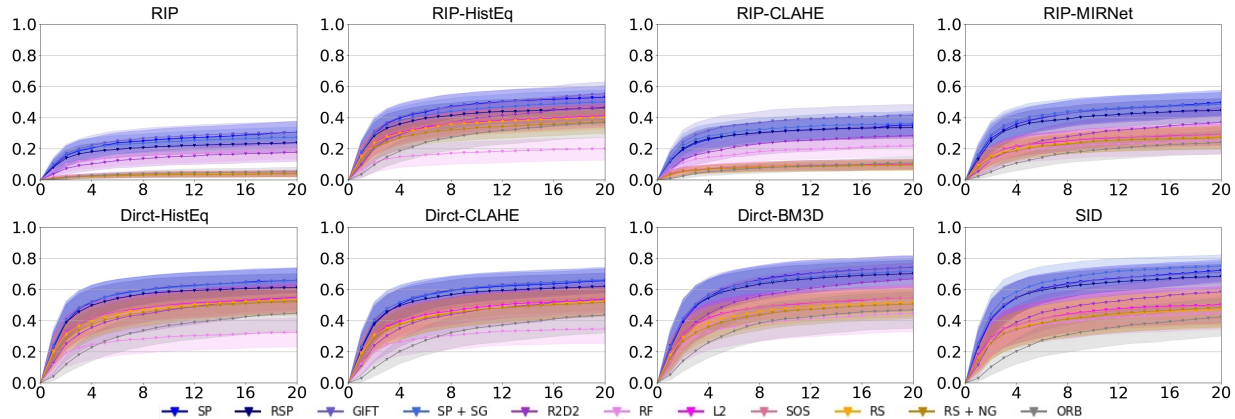
Figure 5. The normalized number N_τ of the exposure settings (the vertical axis) for which the estimation error of each method is lower than threshold τ (the horizontal axis). Each panel shows the means and standard deviations over 54 *indoor* scenes for the eleven image matching methods for an image-enhancing method.

5.2. Results

Figure 5 shows the results for the indoor scenes; see Fig. 11 in the supplementary for the outdoor scenes. Table 1 shows the mean of N_τ with $\tau = 5°$ over 54 scenes for indoor and outdoor scenes, i.e., the values of the curve in Fig. 5 and Fig. 11 at the error threshold $\tau = 5°$. It can be used as a summary of Fig. 5 and Fig. 11. We can make the following observations.

First, the overall comparison of the image enhancers indicates the following: *i) Using the standard RIP to convert RAW-format images to 8-bit RGB images before enhancing and matching is inferior to the direct use of RAW-format images.* This shows that the standard RIP cannot utilize the information stored in the low bits of the RAW signals. This fact forms a basis for our dataset.

Next, the overall comparison of the image matching methods yields the following: *ii) SP and its variants are clearly better than the other methods.* For example, SP and GIFT outperform RS and R2D2 in all cases. This may somewhat contradict previous reports [9, 27] that while SP is superior to SIFT in the homography-based evaluation using the HPatches dataset, the superiority is not observed in the evaluation with non-planar scene matching. Additionally, *iii) SP+SG performs the best in many cases.* However, the gap to other methods considerably differs between the indoor and the outdoor scenes. For the outdoor scenes, the gap to the second-best methods tends to be large, whereas, for the indoor scenes, it is not so large.

The comparison within standard-camera-pipeline-based enhancers indicates the following. *iv) The results of the standard RIP (without any enhancement) are the worst. Comparing RIP-HistEq and RIP-MIRNet, the former is comparable or even better than the latter.* This agrees with the results reported in the recent study of Jenicek and Chum [25], where the authors use 8-bit RGB images outputted

from the standard RIP.

Finally, the comparison within the enhancers using RAW-format images shows the following. *v) For the outdoor scenes, the four enhancers show similar performance in many cases. When used with SP+SG, both BM3D and SID perform better than Direct-HistEq and Direct-CLAHE; the two show the best performance. For the indoor scenes, while there is a similar tendency, SID shows a good margin with others only when used with SP+SG.* It is noteworthy that the superiority of SG depends on the chosen image enhancer, regardless of whether they are applied to indoor or outdoor scenes; this tendency cannot be predicted sorely from the performance of SP.

We conclude that if we use SG, we should choose SID for the image enhancer, which achieves the best performance; if we do not, we should use BM3D since it achieves good performance overall. This conclusion differs from that with the standard-camera-pipeline-based enhancers (i.e., (iv)), which is another evidence that the proposed dataset offers what is unavailable in the previous datasets providing only low-bit depth images. Figure 6 shows the visualization of a few matching results for an indoor scene.

6. Summary and Discussion

This paper has presented a dataset created for evaluating image matching methods for low-light scene images. It contains stereo images of diverse low-light scenes (54 indoor and 54 outdoor scenes). They are captured with 48 different exposure settings, including from mildly to severely underexposed ones. The dataset provides ground truth camera poses to evaluate image matching methods in terms of the accuracy of estimated camera poses.

We have reported the experiments we conducted to test multiple combinations of existing image-enhancing methods and image-matching methods. The results can be sum-

Table 1. Averaged number N_τ over 54 scenes of exposure settings for which each method yields a better result than error threshold $= 5°$. Extracted from Fig. 5 and Fig. 11 in the supplementary. 'R-' means 'RIP-' and 'D-' means 'Direct-.'

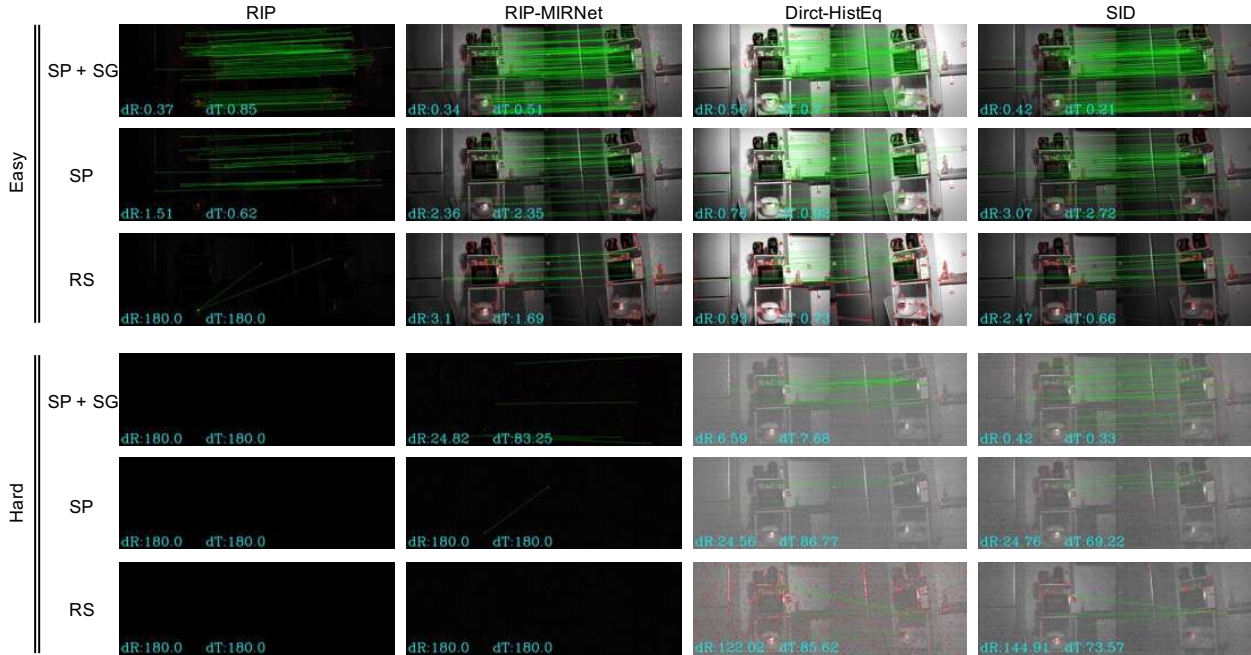|	Indoor								Outdoor							
	RIP	R-HistEq	R-CLAHE	R-MIRNet	D-HistEq	D-CLAHE	D-BM3D	SID	RIP	R-HistEq	R-CLAHE	R-MIRNet	D-HistEq	D-CLAHE	D-BM3D	SID
SP	0.223	0.421	0.275	0.381	0.548	0.540	0.596	0.583	0.233	0.379	0.269	0.352	0.460	0.475	0.502	0.500
RSP	0.190	0.379	0.277	0.365	0.523	0.523	0.581	0.577	0.215	0.363	0.277	0.335	0.435	0.448	0.494	0.477
GIFT	__0.238__	__0.427__	__0.338__	0.390	__0.552__	__0.550__	__0.602__	0.583	0.254	0.375	0.321	0.358	0.475	0.477	0.506	0.492
SP + SG	0.219	0.400	0.292	__0.404__	0.548	0.544	0.585	0.619	__0.302__	__0.419__	__0.365__	__0.410__	__0.525__	__0.527__	0.577	__0.575__
R2D2	0.113	0.317	0.192	0.229	0.388	0.383	0.483	0.421	0.104	0.240	0.163	0.188	0.267	0.277	0.373	0.321
RF	0.138	0.154	0.152	0.192	0.256	0.275	0.346	0.358	0.160	0.146	0.183	0.202	0.225	0.244	0.323	0.325
L2	0.027	0.323	0.077	0.227	0.442	0.415	0.444	0.394	0.052	0.331	0.096	0.258	0.410	0.423	0.427	0.406
SOS	0.029	0.333	0.077	0.229	0.438	0.429	0.440	0.392	0.054	0.325	0.096	0.256	0.417	0.413	0.423	0.402
RS	0.025	0.317	0.071	0.210	0.423	0.404	0.410	0.369	0.046	0.317	0.094	0.242	0.410	0.413	0.404	0.406
RS + NG	0.023	0.288	0.073	0.202	0.404	0.398	0.388	0.363	0.048	0.296	0.102	0.229	0.388	0.392	0.396	0.375
ORB	0.029	0.210	0.056	0.125	0.267	0.238	0.296	0.238	0.069	0.213	0.094	0.144	0.265	0.233	0.277	0.217

Figure 6. Visualization of the matching results for one of the 54 indoor scenes. Point correspondences judged as inliers are shown in green lines. The combination of three matching methods and the four image enhancing methods are applied to two image pairs with different levels of exposure (i.e., 'Easy' and 'Hard').

marized as follows.

- The direct use of the RAW-format images shows a clear advantage over the standard RIP. Using the standard RIP yields only suboptimal performance, as it cannot utilize information stored in the lower bits of RAW-format signals. Moreover, when using the standard RIP, using classical histogram equalization or the state-of-the-art CNN-based image-enhancing method does not make a big difference, as reported in [25].

- SuperPoint and its variants work consistently better than RootSIFT.

- SID is the best image enhancer when using Super-Point+SuperGlue. Otherwise, BM3D and SID perform equally well and better than the sole use of histogram equalization.

While the above is our conclusion about the combinations of currently available methods, we think there remains much room for improvement. For instance, we manually chose the range of 14-bit RAW signal and converted it into 8-bit images, and applied Superpoint to them. It is observed that the manual method yields significantly better results than the image enhancers tested in this paper, showing that none of the tested methods can choose the best range in the 14-bit RAW signals for image matching; see Sec. B in the supplementary for details. The standard image enhancers are designed to yield images that appear the most natural, which should differ from the best image for image matching. We will explore this possibility in a future study.

References

[1] H. Aanæs, A. L. Dahl, and K. S. Pedersen. Interesting interest points. *IJCV*, 97(1):18–35, 2012. 2, 3

[2] P. F. Alcantarilla and T. Solutions. Fast explicit diffusion for accelerated features in nonlinear scale spaces. *IEEE Trans. Pattern Anal. Mach. Intell.*, 34(7):1281–1298, 2011. 1

[3] R. Arandjelović and A. Zisserman. Three things everyone should know to improve object retrieval. In *Proc. CVPR*, 2012. 4, 6

[4] V. Balntas. Silda: A multi-task dataset for evaluating visual localization, 2018. https://research.scape.io/silda/. 2

[5] V. Balntas, E. Johns, L. Tang, and K. Mikolajczyk. Pn-net: Conjoined triple deep network for learning local image descriptors. *arXiv:1601.05030*, 2016. 2, 6

[6] V. Balntas, K. Lenc, A. Vedaldi, and K. Mikolajczyk. Hpatches: A benchmark and evaluation of handcrafted and learned local descriptors. In *Proc. CVPR*, 2017. 2

[7] A. Barroso-Laguna, E. Riba, D. Ponsa, and K. Mikolajczyk. Key. net: Keypoint detection by handcrafted and learned cnn filters. In *Proc. ICCV*, 2019. 2

[8] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. In *Proc. ECCV*, 2006. 2

[9] A. Bhowmik, S. Gumhold, C. Rother, and E. Brachmann. Reinforced feature points: Optimizing feature detection and description for a high-level task. In *Proc. CVPR*, 2020. https://github.com/aritra0593/Reinforced-Feature-Points. 6, 7

[10] J. Bian, W. Lin, Y. Matsushita, S. Yeung, T. Nguyen, and M. Cheng. Gms: Grid-based motion statistics for fast, ultra-robust feature correspondence. In *Proc. CVPR*, 2017. 3

[11] J. Bian, Y. Wu, J. Zhao, Y. Liu, L. Zhang, M. Cheng, and I. Reid. An evaluation of feature matchers for fundamental matrix estimation. *arXiv preprint arXiv:1908.09474*, 2019. 4

[12] E. Brachmann and C. Rother. Neural-guided ransac: Learning where to sample model hypotheses. In *Proc. ICCV*, 2019. https://github.com/vislearn/ngransac. 2, 4, 6

[13] C. Chen, Q. Chen, M. N. Do, and V. Koltun. Seeing motion in the dark. In *Proc. ICCV*, 2019. 2, 3

[14] C. Chen, Q. Chen, J. Xu, and V. Koltun. Learning to see in the dark. In *Proc. CVPR*, 2018. https://github.com/cchen156/Learning-to-See-in-the-Dark. 1, 2, 3, 4, 5

[15] A. Crivellaro, M. Rad, Y. Verdie, K. Moo Yi, P. Fua, and V. Lepetit. Robust 3d object tracking from monocular images using stable parts. *IEEE Trans. Pattern Anal. Mach. Intell.*, 40(6):1465–1479, 2017. 3

[16] K. Dabov, A. Foi, V. Katkovnik, and K. Egiazarian. Image denoising by sparse 3-d transform-domain collaborative filtering. *IEEE Trans. Image Process.*, 16(8):2080–2095, 2007. 2, 5

[17] D. DeTone, T. Malisiewicz, and A. Rabinovich. Superpoint: Self-supervised interest point detection and description. In *Proc. CVPRW*, 2018. https://github.com/magicleap/SuperPointPretrainedNetwork. 2, 6

[18] X. Dong, G. Wang, Y. Pang, W. Li, J. Wen, W. Meng, and Y. Lu. Fast efficient algorithm for enhancement of low lighting video. In *Proc. ICME*, 2011. 3

[19] M. Dusmanu, I. Rocco, T. Pajdla, M. Pollefeys, J. Sivic, A. Torii, and T. Sattler. D2-net: A trainable cnn for joint detection and description of local features. In *Proc. CVPR*, 2019. 2

[20] M. A. Fischler and R. C. Bolles. Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981. 2

[21] A. Geiger, P. Lenz, and R. Urtasun. Are we ready for autonomous driving? the kitti vision benchmark suite. In *Proc. CVPR*, 2012. 2

[22] X. Guo, Y. Li, and H. Ling. Lime: Low-light image enhancement via illumination map estimation. *IEEE Trans. Image Process.*, 26(2):982–993, 2016. 3

[23] X. Han, T. Leung, Y. Jia, R. Sukthankar, and A. C. Berg. Matchnet: Unifying feature and metric learning for patch-based matching. In *Proc. CVPR*, 2015. 2, 6

[24] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003. 1

[25] T. Jenicek and O. Chum. No fear of the dark: Image retrieval under varying illumination conditions. In *Proc. ICCV*, 2019. 2, 7, 8

[26] R. Jerome, W. Philippe, R. S. César, and H. Martin. R2D2: repeatable and reliable detector and descriptor. In *Proc. NeurIPS*, 2019. https://github.com/naver/r2d2. 2, 6

[27] Y. Jin, D. Mishkin, A. Mishchuk, J. Matas, P. Fua, K. M. Yi, and E. Trulls. Image matching across wide baselines: From paper to practice. *IJCV*, 2020. 2, 3, 7

[28] L. Karel, G. Varun, and V. Andrea. Vlbenchmarks, 2011. http://www.vlfeat.org/benchmarks/. 2

[29] X. Liu, M. Suganuma, Z. Sun, and T. Okatani. Dual residual networks leveraging the potential of paired operations for image restoration. In *Proc. CVPR*, 2019. 2

[30] Y. Liu, Z. Shen, Z. Lin, S. Peng, H. Bao, and X. Zhou. Gift: Learning transformation-invariant dense visual descriptors via group cnns. In *Proc. NeurIPS*, 2019. https://github.com/zju3dv/GIFT. 6

[31] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *IJCV*, 60(2):91–110, 2004. 1, 2, 4, 6

[32] A. Łoza, D. R. Bull, P. R. Hill, and A. M. Achim. Automatic contrast enhancement of low-light images based on local statistics of wavelet coefficients. *Digital Signal Processing*, 23(6):1856–1866, 2013. 3

[33] W. Maddern, G. Pascoe, C. Linegar, and P. Newman. 1 year, 1000 km: The oxford robotcar dataset. *International Journal of Robotics Research*, 36(1):3–15, 2017. 2

[34] H. Malm, M. Oskarsson, E. Warrant, P. Clarberg, J. Hasselgren, and C. Lejdfors. Adaptive enhancement and noise reduction in very low light-level video. In *Proc. ICCV*, 2007. 3

[35] K. Mikolajczyk and C. Schmid. Scale & affine invariant interest point detectors. *IJCV*, 60(1):63–86, 2004. 1

[36] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(10):1615–1630, 2005. 2, 3

[37] K. Moo Yi, E. Trulls, Y. Ono, V. Lepetit, M. Salzmann, and P. Fua. Learning to find good correspondences. In *Proc. CVPR*, 2018. 2, 6

[38] R. Mur-Artal, J. Maria M. Montiel, and J. D. Tardos. Orb-slam: a versatile and accurate monocular slam system. *IEEE Trans. Robotics*, 31(5):1147–1163, 2015. 1

[39] H. Noh, A. Araujo, J. Sim, T. Weyand, and B. Han. Large-scale image retrieval with attentive deep local features. In *Proc. ICCV*, 2017. 2

[40] Y. Ono, E. Trulls, P. Fua, and K. M. Yi. Lf-net: learning local features from images. In *Proc. NeurIPS*, 2018. 2

[41] S. Park, S. Yu, B. Moon, S. Ko, and J. Paik. Low-light image enhancement using variational optimization-based retinex model. *IEEE Trans. Consumer Electronics*, 63(2):178–184, 2017. 3

[42] M. Pultar, D. Mishkin, and J. Matas. Leveraging outdoor webcams for local descriptor learning. In *Proc. Computer Vision Winter Workshop*, 2019. 2

[43] R. Raguram, J. Frahm, and M. Pollefeys. A comparative analysis of ransac techniques leading to adaptive real-time random sample consensus. In *Proc. ECCV*, 2008. 2

[44] E. Rublee, V. Rabaud, K. Konolige, and G. Bradski. Orb: An efficient alternative to sift or surf. In *Proc. ICCV*, 2011. 1, 6

[45] P. Sarlin, D. DeTone, T. Malisiewicz, and A. Rabinovich. Superglue: Learning feature matching with graph neural networks. In *Proc. CVPR*, 2020. https://github.com/magicleap/SuperGluePretrainedNetwork. 2, 6

[46] T. Sattler, W. Maddern, C. Toft, A. Torii, L. Hammarstrand, E. Stenborg, D. Safari, M. Okutomi, M. Pollefeys, J. Sivic, et al. Benchmarking 6dof outdoor visual localization in changing conditions. In *Proc. CVPR*, 2018. 2

[47] T. Sattler, T. Weyand, B. Leibe, and L. Kobbelt. Image retrieval for image-based localization revisited. In *Proc. BMVC*, 2012. 2

[48] N. Savinov, A. Seki, L. Ladicky, T. Sattler, and M. Pollefeys. Quad-networks: unsupervised learning to rank for interest point detection. In *Proc. CVPR*, 2017. 2

[49] J. L. Schonberger, H. Hardmeier, T. Sattler, and M. Pollefeys. Comparative evaluation of hand-crafted and learned local features. In *Proc. CVPR*, 2017. 3

[50] X. Shen, C. Wang, X. Li, Z. Yu, J. Li, C. Wen, M. Cheng, and Z. He. Rf-net: An end-to-end image matching network based on receptive field. In *Proc. CVPR*, 2019. https://github.com/Xylon-Sean/rfnet. 6

[51] E. Simo-Serra, E. Trulls, L. Ferraz, I. Kokkinos, P. Fua, and F. Moreno-Noguer. Discriminative learning of deep convolutional feature point descriptors. In *Proc. ICCV*, 2015. 2, 6

[52] C. Strecha, W. Von Hansen, L. Van Gool, P. Fua, and U. Thoennessen. On benchmarking camera calibration and multi-view stereo for high resolution imagery. In *Proc. CVPR*, 2008. 2

[53] K. Tateno, F. Tombari, I. Laina, and N. Navab. Cnn-slam: Real-time dense monocular slam with learned depth prediction. In *Proc. CVPR*, 2017. 1

[54] Y. Tian, B. Fan, and F. Wu. L2-net: Deep learning of discriminative patch descriptor in euclidean space. In *Proc. CVPR*, 2017. https://github.com/ubc-vision/image-matching-benchmark. 6

[55] Y. Tian, X. Yu, B. Fan, F. Wu, H. Heijnen, and V. Balntas. Sosnet: Second order similarity regularization for local descriptor learning. In *Proc. CVPR*, 2019. https://github.com/ubc-vision/image-matching-benchmark. 6

[56] Y. Verdie, K. Yi, P. Fua, and V. Lepetit. Tilde: A temporally invariant learned detector. In *Proc. CVPR*, 2015. 2, 6

[57] K. Wei, Y. Fu, J. Yang, and H. Huang. A physics-based noise formation model for extreme low-light raw denoising. In *Proc. CVPR*, 2020. 2, 3

[58] X. Wei, Y. Zhang, Z. Li, Y. Fu, and X. Xue. Deepsfm: Structure from motion via deep bundle adjustment. In *Proc. ECCV*, 2020. 1

[59] K. M. Yi, E. Trulls, V. Lepetit, and P. Fua. Lift: Learned invariant feature transform. In *Proc. ECCV*, 2016. 2

[60] D. Yoo, S. Park, J. Lee, and I. S. Kweon. Multi-scale pyramid pooling for deep convolutional representation. In *Proc. CVPRW*, 2015. 2, 6

[61] S. Zagoruyko and N. Komodakis. Learning to compare image patches via convolutional neural networks. In *Proc. CVPR*, 2015. 2

[62] S. W. Zamir, A. Arora, S. Khan, M. Hayat, F. S. Khan, M.-H. Yang, and L. Shao. Learning enriched features for real image restoration and enhancement. *arXiv preprint arXiv:2003.06792*, 2020. 2, 5

[63] Y. Zhang, K. Li, K. Li, B. Zhong, and Y. Fu. Residual non-local attention networks for image restoration. In *Proc. ICLR*, 2019. 2

[64] C. L. Zitnick and K. Ramnath. Edge foci interest points. In *Proc. ICCV*, 2011. 2